16

Language Documentation & Conservation Special Publication No. 15
Reflections on Language Documentation 20 Years after Himmelmann 1998
ed. by Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton, pp. 157–172
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24817

Reflections on documentary corpora

Sally Rice
University of Alberta

For decades, language documentation proponents have argued for the separability of LD as its own sub-discipline. Many corpus linguists have made this same claim; thus, corpus linguistics shares the ethos of data over theorizing, whereby primary data represent authentic, connected discourse that is natural (not elicited), broadly sampled (across speakers, generations, dialects), and balanced (reflecting different usage contexts and genres). Nevertheless, many misconceptions remain about what a language corpus is, how it is formatted, how big or balanced it needs to be, and most importantly, *how it is queried*. In this reflection, I dispel some of these misconceptions, while reassuring community members and field linguists alike that a corpus is an exceedingly powerful tool for guiding the expansion of the documentary record, keeping precious language data in circulation, and helping to produce the classic descriptive by-products of LD such as dictionaries, phrasebooks, and grammars. Above all, the less-familiar but more direct by-products of corpus interrogation, such as word lists, frequency counts, concordance lines, N-grams, collocations, distribution, and dispersion plots, are so immediately interpretable and useful by speakers, learners, and linguists, that LD should give corpus linguistic training the same attention as project planning, ethics, recording, transcription, annotation, metadata, and archiving.

1. When documenting "linguistic practices" becomes focusing on actual spoken usage If the purpose of language documentation (LD), as so persuasively argued and succinctly crystallized by Nikolaus Himmelmann (1998: 166), is to provide "a comprehensive record of the linguistic practices characteristic of a speech community", then a corpus is truly an excellent means of achieving this in ways readily accessible to speakers, learners, and outsider linguists. Indeed, Himmelmann wrote of the need to compile a collection or corpus of "communicative events", recognizing, if only tacitly, that LD typically transpires in the context of oralcy; thus, spontaneous interactive conversation should be the centerpiece of documentary efforts. Himmelmann's original articulation two decades ago (echoed and amplified by Woodbury 2003) of how documentary linguistics might especially focus on a different kind of primary data—

ISBN: 978-0-9973295-3-7

connected, naturally-occurring speech, whether narrative or conversation—dovetails more or less with recognition among corpus linguists that spoken language constitutes an equally important and thoroughly different mode of language use than that found in written genres. In the 1980s and 1990s, large national corpora for major languages like English pushed hard to include transcripted samples of spoken varieties alongside more easily compiled textual samples from newspapers, fiction, and academic writing. Insights about the profound differences between spoken and written modalities of language ensued (cf. the magnificent *Longman Grammar of Spoken and Written English*, Biber et al. 1999, based on the Longman Corpus Network corpora described at `www.global.longmandictionaries.com/Longman/corpus`); true corpus-based grammars and dictionaries of multiple languages also followed, as did new varieties of corpora, including learner, parallel, conversational, and multimodal corpora.

Despite the increasing recognition of the role that corpora play in LD and linguistics generally, there remain some entrenched misapprehensions about what a language corpus is and what one can do with such a corpus (be it big or small, balanced or skewed, annotated or not). In this reflection, I applaud the increase in calls for corpus-building in the LD and field linguistics literature (§2), spell out some of the prevailing misconceptions about language corpora in LD circles from the viewpoint of corpus linguistics proper (§3), and put the well documented "front-end" challenges of *building* a corpus in the first place (§4) alongside some of the many "back-end" benefits of *using* a corpus in the second place (§5). (Note, I intend *front-end/back-end* to be meant temporally, not in typical computational parlance of accessible/inaccessible to the user.) Chief among these benefits is getting a broader and sharper picture of actual spoken language usage patterns and patterns of variation within a speech community, a picture that can help inform subsequent stages of documentation.

2. Singing the virtues of documentary corpora: A rising chorus Since the publication of Himmelmann 1998, there has been a steady increase in edited volumes, textbooks, and handbooks about field linguistics and language documentation. Table 1 provides a list of some of the major book-length publications of the past two decades, arranged chronologically and showing the number of pages in the index under the heading *corpus/corpora* and the percent this represents against the total page number in each volume—an admittedly poor metric of attention, given the high degree of variability in indexing specificity and practice.

The notion of building documentary corpora is evidently growing more prevalent in the LD literature; see the steady upwards trend line in Figure 1, which graphically represents the percent frequency of mention of the words *corpus* or *corpora* by page in the volumes listed in Table 1. Sadly, it is still rare to find any listing for *conversation*, *speech*, or *interaction* in the typical LD index—the usual source of the primary data supposedly feeding into documentary corpora.

While it is heartening to see the role of corpora in LD being increasingly recognized (cf. McEnery & Ostler 2000; Scannell 2007; Mosel 2014), problematized (cf. Johnson 2004; Cox 2011; Jung & Himmelmann 2011; Vinogradov 2016), and evaluated (cf. Thieberger et al. 2015; Thieberger 2016), the field has a long way to go in understanding what a corpus is and is not. Moreover, the LD use of the word *corpus* as in *documentary corpus* is quite different from how a corpus linguist views the term. The focus in LD is generally on compiling the corpus, giving short shrift to what to do with the corpus data so compiled.

	Title	Corpus pages	Total pages	%
A	Newman & Ratliff (eds.) (2001). *LF.*	0	288	0%
B	Hinton & Hale (eds.) (2001). *Green Book of LR in Practice.*	1	468	0.2%
C	Gippert, Himmelmann, & Mosel (eds.) (2006). *Essentials of LD.*	47	424	11%
D	Crowley (2007). *FL: A Beginner's Guide.*	3	202	1.5%
E	Bowern (2008). *LF: A Practical Guide.*	8	285	3%
F	Grenoble & Furbee (eds.) (2010). *LD: Practice & Values.*	31	340	9%
G	Austin & Sallabank (eds.) (2011). *Cambridge Handbook of EL.*	38	567	7%
H	Chelliah & de Reuse (2011). *Handbook of Descriptive LF.*	21	492	4%
I	Haig et al. (eds.) (2011). *Documenting EL.*	1	344	0.2%
J	Thieberger (ed.) (2012). *Oxford Handbook of LF.*	38	545	7%
K	Sakel & Everett (2012). *LF: A Student Guide.*	2	179	1%
L	Jones & Ogilvie (eds.) (2013). *Keeping Languages Alive.*	6	269	2%
M	Jones (ed.) (2015). *EL & New Technologies.*	30	211	14%

Table 1: Number of pages in major LD publication indices mentioning corpus/corpora as a percentage of total pages overall. Volumes are listed chronologically. EL=endangered languages; FL=field linguistics; LD=language documentation; LF=linguistic fieldwork; LR=language revitalization.

3. Lingering misconceptions about what a corpus is Since Himmelmann 1998 first distinguished linguistic description and language documentation, the latter has become associated with collecting primary data in the form of audio and video recordings, making transcriptions and other annotations of such recordings, and compiling these transcribed representations, with appropriate metadata, into a corpus for archiving. Himmelmann's own view on using a documentary corpus suggests that it has "at least the potential of being of use to a larger group of interested parties. These include the speech community itself, which might be interested in a record of its linguistic practices and traditions" (ibid.: 163). This is an exceedingly vague and uninspiring illustration of the application of a documentary corpus. Indeed, the bulk of this seminal article is about corpus compilation, from the sampling of a full array of communicative event types to the metadata annotation that primary recordings and secondary transcriptions should receive.

Himmelmann 1998 definitely set the stage for codifying what I'm calling the *front-end* protocols of documentary linguistics: speaker sampling and recording techniques, transcription and annotation, metadata management and archiving. This much-needed attention has continued through Thieberger & Berez 2012 and just about every volume listed in Table 1. Unfortunately, these sources are usually replete with elusive and ultimately off-hand comments that do little to clarify exactly what a corpus is capable of. In several instances, the quotes in (1) exhaust the topic of *corpus* in their respective

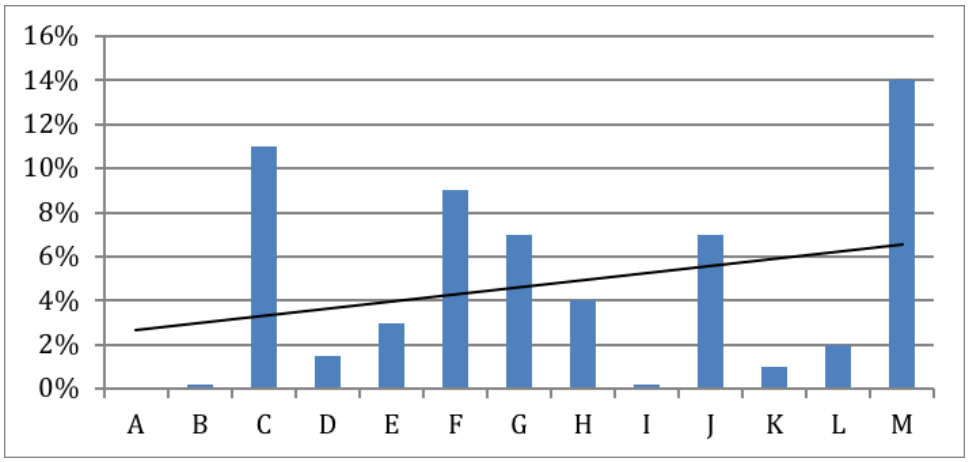


Figure 1: Slight but steady increase (see trend line) in percentage of pages in major LD publication indices mentioning corpus/corpora from 2001-2015, as listed chronologically and described in Table 1.

sources. Critically, they lead nowhere that an uninformed fieldworker never exposed to corpus linguistics can follow.

(1) Some minimalist comments about *using* a corpus in the LD literature

a. "Corpus data is more useful if it's annotated. That allows you to search for more detailed environments. It also allows you to create sub-corpora…that would let you search for differences between…two genres." (Bowern 2008: 120);

b. "Even given a large corpus of data, we may not have enough information to interpret the data without analysis. Thus, it is difficult to know whether all the linguistic forms and structures have been represented by the available data, whether paradigmatic gaps are intentional or rare, and what types of linguistic elicitations are needed to fill out the corpus of data." (Berge 2010: 54);

c. "The corpus should be annotated in a way that would allow a philologist in the distant future to interpret its content." (Good 2010: 126);

d. "There is absolutely no reason why the kinds of corpus-based statistical studies that have been carried out extensively on different varieties of English could not be carried out in other languages as well." (Crowley 2007: 18);

e. "Corpus linguistics does not typically result from the activities of fieldworkers, since corpora typically consist of written data easily studied by computational methods, although they are increasingly transcripts from spoken data." (Chelliah & de Reuse 2011:12);

f. "A well-formed corpus allows us to seek answers to linguistic questions that are difficult to ask when data is limited to what can be expressed on the printed page." (Thieberger & Berez 2012: 116).

Indeed, in otherwise excellent overviews of creating and annotating language corpora, Vinogradov (2016) and Gries & Berez (2017) compare some basic characteristics of classic by-products of LD such as a Boasian text collection in terms of a variety of features, as shown in Table 2.

I have highlighted the last two features, *searchability* and *quantitative analysis*, because both are left as casually referenced and unexplained as the activities listed above in (1). Any corpus, however large or small, affords a birds' eye view of the material therein. It is this ability to search materials in the aggregate that allows the emergence of language-specific patterns that go beyond the anecdotal. Indeed, depending on the size of the corpus, some observations about pattern frequency can be statistically confirmed through simple association measures or openly challenged with more data. These patterns may be very fragmentary and low-level, but as recurrent expressions they generally constitute the core of actual language-in-use.

The heart of the matter is this: Suppose you're a middle-aged (or older) field linguist who came of age before the emergence of corpus linguistics or suppose you're an undergraduate or graduate student being trained in LD at a university that doesn't offer corpus linguistics training (which still describes the majority of linguistics departments)? How are you to square the circle between corpus creation and corpus application if you have never worked with a concordancer (the generic name for corpus-querying software), never queried multiple corpus files at the same time, never found strange patterns of co- or non-occurrence, never been surprised by the large number of fixed expressions that turn up, or never really confronted the staggering differences in frequency between lexical and grammatical material in a language or the idiosyncratic distribution of particular words or phrases in different genre types? Understanding what a corpus is and what it can do is only going to enhance and motivate the LD process itself. Going forward, we must stop regarding the corpus as a body of recordings, impeccably textualized and identified, and possibly left silent and still in an archive, but instead view it as an active and noisy collection of transcribed conversations teeming with insights about the language and its use that we can eavesdrop on again and again.

4. What a language corpus—documentary or otherwise—really is A corpus is neither a field linguistic database (as in a FLEx-style project with elicited fieldnotes, interlinearized utterances or narratives, a morpheme and word lexicon, etc.) nor a text collection. At its most basic, a corpus is a machine-readable collection of text *files* that can be queried simultaneously or selectively. In the case of spoken corpora—as documentary corpora are most likely to be—those digital text files will consist of transcriptions of speech (the output of transcription software such as ELAN, which allows for time-aligned annotation of an audio/video signal). If the speech source reflects unplanned conversation, then there will likely be incomplete utterances, repetitions, hesitations, interruptions, over-speech, all segmented into turns or intonation units. If the speech source reflects more planned narrative (a personal story, traditional legend, or oratory), then the transcribed text file may evidence more holistic, sentence-like structures. Regardless, both broad types of spoken language share the virtue of being natural and contextualized. Together with other communicative event types, they can form a corpus of mono- and dialogic language use as recorded in a speech community. Since the transcription (text) files are backed up by media as well as copious metadata, they themselves need not and should not contain any other information beyond the

Feature	Major corpora	Documentary corpora	Language archives	Printed text collection
selectivity of material	+	+	–	+/–
machine-readable format	+	+	+	–
volume	big/huge	small	big/small	very small
annotation	+	+/–	–/+	+/–
balanced subcorpora	+	–/+	–/+	–
searchability	+	+	–/+	–
quantitative analysis	+	–/+	–	–

Table 2: Presence or absence of basic characteristics of different research instruments for LD (adapted from Vinogradov 2016: 136 (his Table 3) and Gries & Berez 2017.

transcription and possibly an identifier for the speaker at each interactional turn. The text files that constitute the language corpus are not the same as the transcription files.

Here, I switch to a new moniker, language corpus or LC, to distinguish it from the documentary corpus, DC, that not only means something else, but is too often associated with inexact and promissory applications. A DC is about collecting and cataloguing data. An LC is about effectively and imaginatively exploring those data. Any time-stamped transcriptions, which may be further parsed, interlinearized, tagged, lemmatized, and translated into another language with attendant situational metadata, belong in the archived DC. Ultimately, the LC should be monolingual (code-switching aside). It can also be small, unbalanced, un-annotated and un-lemmatized (no reduction of inflected or derived forms to their bare stem). This lack of mark-up beyond a clean and consistent transcription is especially relevant in the earliest stages of LD when data are scarce, analytical knowledge is lacking, and the time and energy to annotate are in short supply (cf. Boerger 2011). Whereas these limitations can cripple language description and analysis, they constitute virtues in certain corpus linguistic camps, such as the neo-Firthians or the Birmingham School (cf. Sinclair 1991 and, especially, McEnery & Hardie 2012, Chapter 6, for helpful overviews), which regard corpus returns of un-annotated text or speech samples rather than linguistic theory or typological/areal expectation as the ultimate arbiter of what's going on in a language.

A real LC is not an archive of available material. It involves the rendering of that material to be machine-readable and query-able. In short, the LC is a folder, stratified or not, composed of a set of appropriately named text files. These files should have transparent file names that identify attributes deemed relevant to the particular LD project (e.g. speaker ID, genre, dialect, recording date, link to media file, etc.). Concordancers return data from queries linked to their source files, so good file-naming (the only place that metadata should reside) is especially pertinent during actual corpus searches.

5. What a language corpus can do (the neglected back-end) Thus far, I've lamented how applications of a corpus are left implicit in much of the LD literature. It's now time to be explicit and put a sample demonstration corpus through its paces. In (2) and (3), I list some common concordancer tools and corpus linguistic applications. The screen shots illustrated in Figures 2–7 are taken from Rice & Thunder 2017 and reflect data from a nearly 9,000-word corpus of *nêhiyawêwin* (or Plains Cree; ISO 639-3: cre), an Indigenous language of Western Canada, comprised of nine files representing three genres: casual conversation (C), planned narrative (N), and written stories (S). The demonstration concordancer into which the nine files were uploaded is AntConc (Anthony 2018), which can handle UTF-8 encoded (Unicode) plain text files and even help identify inconsistencies in spelling or file-rendering when files are first uploaded.

(2) Some classic concordancer tools

   a. orthographic or frequency-based word lists, as in Figure 2;
   b. keywords-in-context (KWIC), also known as concordance lines, as in Figure 3;
   c. N-grams or recurrent fixed expressions of various lengths, as in Figure 4;
   d. collocates of an item, be it morpheme, word, or expression, as in Figure 5;
   e. dispersion plots (which locate where in a file a certain string, be it morpheme, word, or phrase appears), as in Figure 6;

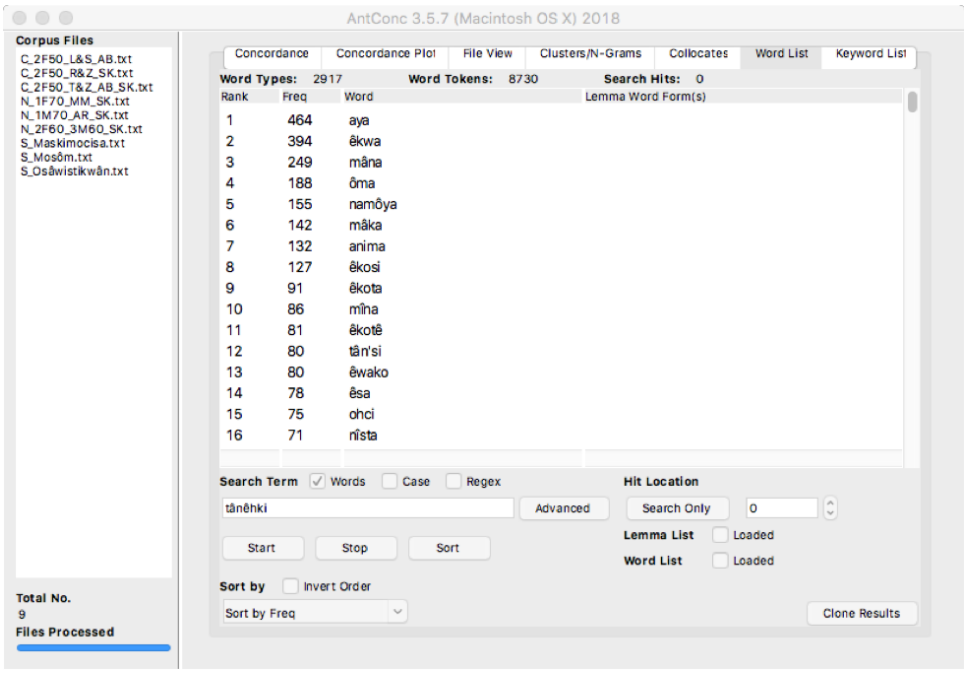


Figure 2: A frequency-based word list returned from the *nêhiyawêwin* demonstration corpus using the AntConc **Word List** function. Knowing which words are highly recurrent versus rare or absent in a corpus or in particular corpus files helps both the linguist and the language instructor target phenomena to investigate or teach. Here, the top-ranked word *aya* is a hesitation device. The next most frequent item *êkwa* 'and' is a conjunction.

f. regular expression (regex) searches, using wildcards and other simple scripts to search within or across words, as in Figure 7.

(3) Some classic corpus applications

a. build exemplified dictionaries and grammars by providing a source of natural, example sentences (via concordance lines);

b. help with synonymy differentiation;

c. allow for sense disambiguation;

d. provide context for discoveries about semantic prosody;

e. demonstrate genre/register/dialect/gender/generational differences;

f. identify useful recurrent expressions, formulaic language, or phrasemes (as discussed in Rice 2017) that can help learners begin to develop conversational skills.

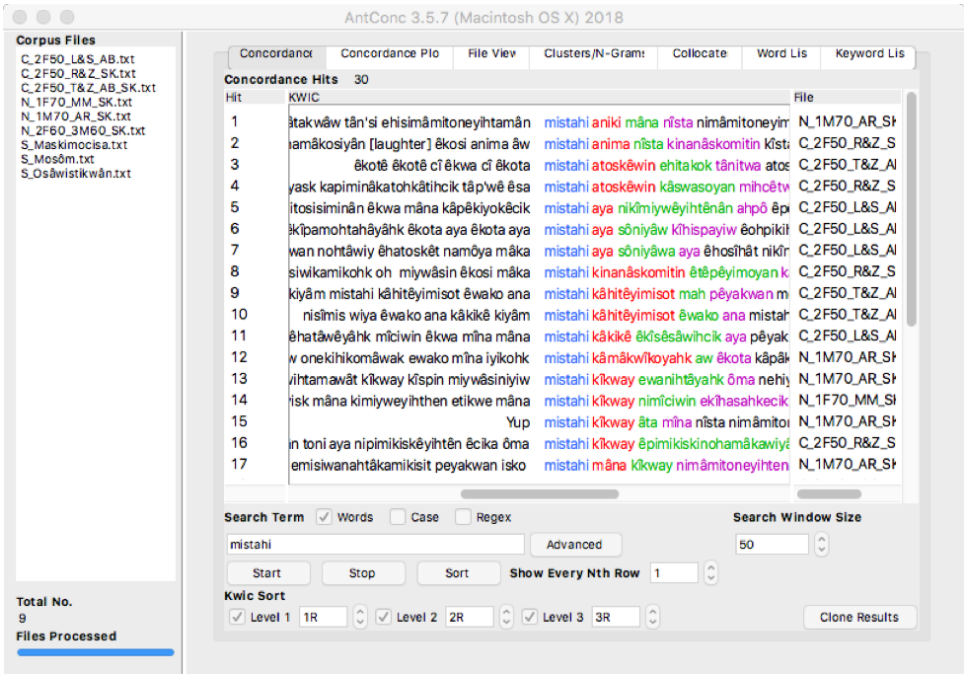


Figure 3: A set of concordance lines returned from a search of *mistahi* 'a lot of' sorted by first, second, and third word to the right using the Concordance function. From Hit lines 13-16, we can see that the phrase, *mistahi kîkway* 'a lot of something', appears four times in the corpus across four distinct files: three narratives and one conversation. If one knew nothing about the language, the prevalence of this bigram in such a small corpus would suggest that it has some sort of unit status as an expression.

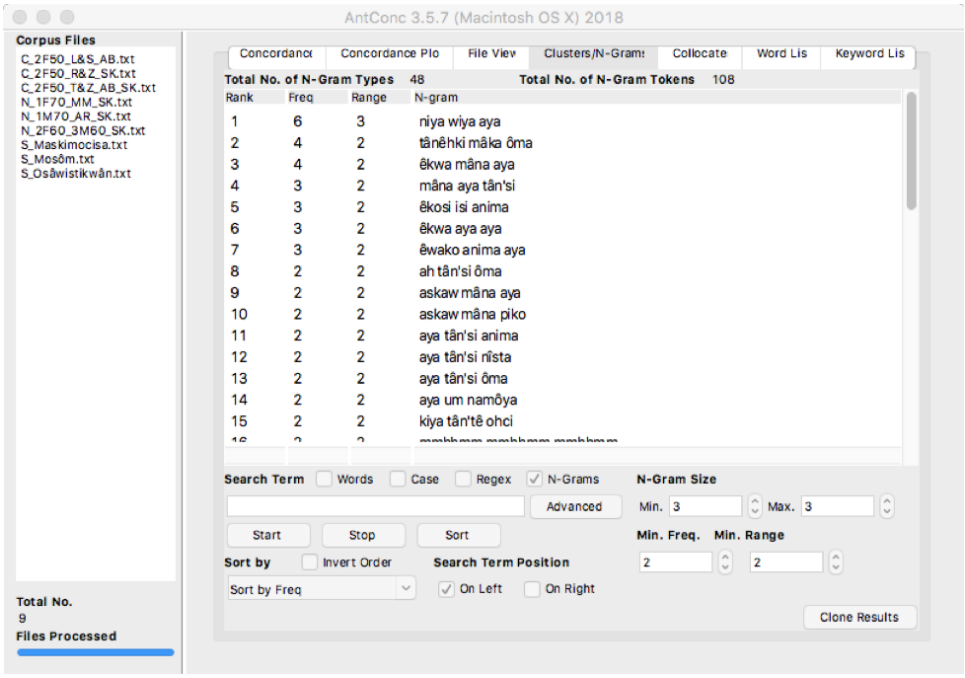


Figure 4: A set of 3-grams with a frequency of at least 2 and a range (number of files) of at least 2 returned using the **Clusters/N-Grams** function. This function can indeed launch a fishing expedition. We are asking the corpus to look for patterns of three recurrent words without any preconception as to their meaning or structure. In this case, of the 15 visible returns in the list, *aya* (a hesitation device), surfaces in 10 or $2/3^{rds}$ of the cases. In text or prepared narrative, any recurrent multi-word strings would likely be more informative and point to actual fixed expressions in the language.

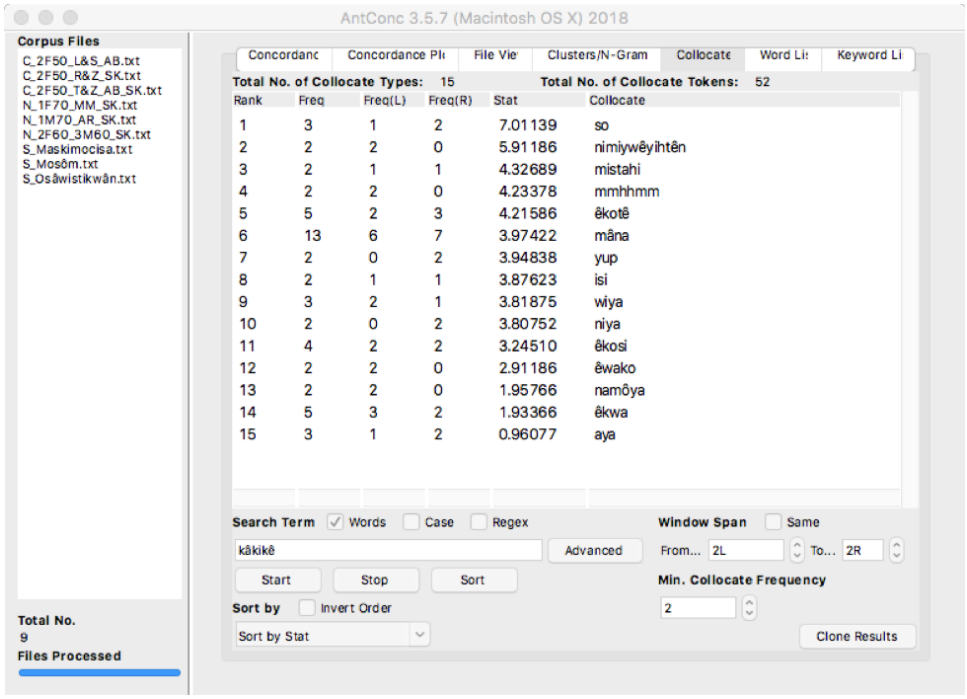
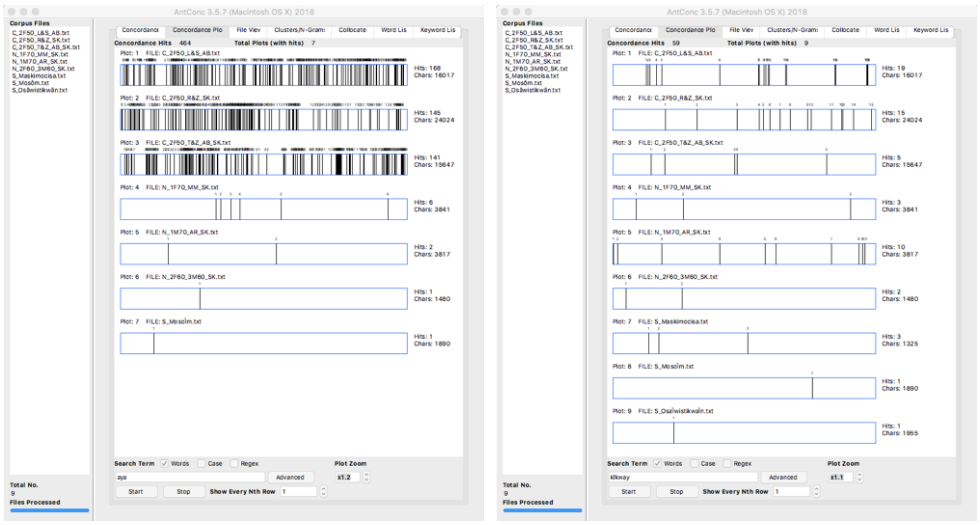


Figure 5: Collocates of *kâkikê* 'always' within 2 words to the left or right with a frequency of at least 2 and a range of at least 2 returned using the **Collocate** function. This function gives an indication of words that tend to co-occur within a fixed span, even though they might not be adjacent, such as *very* and *indeed* in varieties of English which frequently surface with an intervening adjective or adjectival phrase of varying length.



(a) **(b)**

Figure 6: Two sets of dispersion plots for the hesitation device, *aya*, in (a) and the indefinite pronoun, *kîkway*, in (b) showing, respectively, the highly skewed or relatively well distributed occurrence of each item within each corpus file. These results were returned using the **Concordance Plot** function. This function can yield immediate insights into differences in genre, speaker, etc., as well as differences in words that have more of a lexical vs. more of a grammatical function in the language.

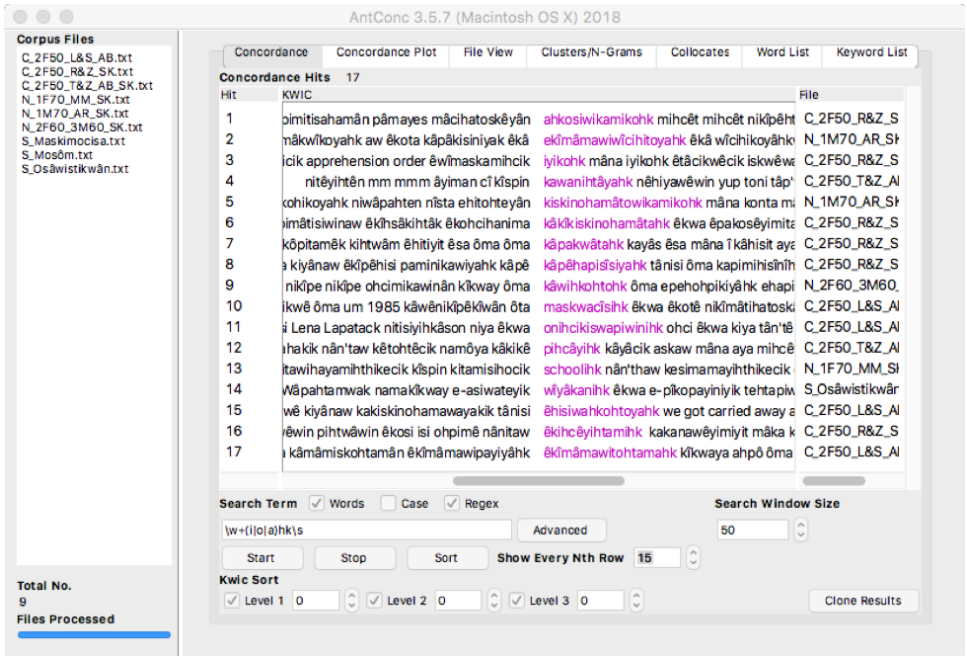


Figure 7: Concordance lines returned from a regex search using the **Concordance** function. Although there are three allomorphs in *nêhiyawêwin* of the locative suffix, {*-ihk*, *-ohk*, *-ahk*}, words ending in all three variants can be queried simultaneously using a regular expression such as \w+(i|o|a)hk\s. The use of regular expressions when conducting corpus searches helps overcome challenges caused by allomorphy, variation in spelling, incomplete knowledge, or other context effects that may affect a form.

Two widely subscribed LD maxims are also shared by corpus linguists: (i) taking a language as it comes (not based on translation, elicitation, or someone else's analysis) and (ii) making samples of language accessible and re-useable for multiple purposes and users. If we must all do as much as we can with the language samples we've got, then the multiple queries that can be conducted on a language corpus by a concordancer seem downright economical and efficient. There is huge bang for the corpus buck, in both early and late stages of LD. Seeing data displayed in the form of corpus returns also serves as an inspiration and a directive to collect more samples more broadly from more usage situations and speakers, if at all possible. A small, untagged, and unbalanced corpus can still yield tremendous insights into the structure, meaning, and use of a language—sampling skews never go away, regardless of corpus size. Most endangered language communities or LD projects led by a single individual probably have all the tools and personnel needed to start building and using a language corpus. The creation and maintenance of such a corpus can involve a variety of community members with differing skills and interests, from recording and transcription to file-editing and metadata management (cf. Boerger 2011). Community-led, corpus-based LD projects can go hand-in-hand without much or any intervention from a linguist or programmer. Amongst the many new skill sets that field linguists and endangered language activists need to develop—beyond linguistic analysis, ethical conduct, grant-writing, and front-end protocols—should be a basic understanding of corpus linguistics.

In re-conceptualizing the documentary corpus as an actualized, query-able corpus of everyday conversation or communicative events, the benefits of corpus-creation and the bounties afforded by interrogating such a corpus with proper concordancing tools can be explicitly demonstrated, demystified, and hopefully implemented widely by speakers and learners in endangered and minority language speech communities. With the availability of free, off-the-shelf, easy-to-use, Unicode-savvy, XML-capable, multi-platform, 4th generation (stand-alone) concordancers such as AntConc, a corpus does not have to live on-line, but can reside on a computer (or two) in a community. Thus, LD and documentary corpora will be able to achieve a few of the widely held desiderata itemized by Bird & Simons 2003, Woodbury 2003, Himmelmann 2006, and others: a lasting, multipurpose, and re-useable product of documentary efforts. The LD field has spent enough time talking about coverage. It's time to leverage that coverage into actually applying corpus tools and conducting corpus analyses that allow precious language data to speak for themselves without descriptive or analytic overlay and, most importantly, without further delay.

References

Anthony, Laurence. 2018. AntConc: A freeware concordance program for Windows, Macintosh OSX, and Linux (Version 3.5.6) [Computer Software]. Tokyo: Waseda University. Available from `http://www.laurenceanthony.net/`.

Austin, Peter K. & Julia Sallabank (eds.). 2011. *The Cambridge handbook of endangered languages.* Cambridge: Cambridge University Press.

Berge, Anna. 2010. Adequacy in documentation. In Grenoble & Furbee (eds.), *Language documentation: Practice and values*, 51–66. Amsterdam/New York: John Benjamins.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English.* London: Longman.

Bird, Steven & Simons, Gary. 2003. Seven dimensions of portability for language documentation. *Language* 79(3). 557–582.

Boerger, Brenda. 2011. BOLDly go where no one has gone before. *Language Documentation & Conservation* 5. 208–233.

Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*, 1st edn. London/NY: Palgrave MacMillan.

Chelliah, Shobhana & Willem de Reuse. 2011. *Handbook of descriptive linguistic fieldwork.* Dordrecht: Springer.

Cox, Christopher. 2011. Corpus linguistics and language documentation: Challenges for collaboration. In John Newman, R. Harald Baayen & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, 239–264. Amsterdam: Brill.

Crowley, Terry. 2007. *Field linguistics: A beginner's guide.* Oxford: Oxford University Press.

Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.). 2006. *Essentials of language documentation.* Berlin: Mouton de Gruyter.

Good, Jeff. 2010. Valuing technology: Finding the linguist's place in a new technological universe. In Grenoble & Furbee (eds.), *Language documentation: Practice and values*, 111–131. Amsterdam/New York: John Benjamins.

Grenoble, Lenore A., N. Louanna Furbee (eds.). 2010. *Language documentation: Practice and values.* Amsterdam/New York: John Benjamins.

Gries, Stefan Th. & Andrea Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (eds.), *Handbook of linguistic annotation*, 379–409. Dordrecht: Springer.

Haig, Geoffrey, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.). 2011. *Documenting endangered languages: Achievements and perspectives.* Berlin: de Gruyter.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161-195.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin: Mouton de Gruyter.

Hinton, Leanne & Hale, Ken (eds.). 2001. *The green book of language revitalization in practice.* New York: Academic Press.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.), *Language documentation and description*, vol. 2, 140–153. London: SOAS.

Jones, Mari C. (ed.). 2015. *Endangered languages and new technologies*. Cambridge: Cambridge University Press.

Jones, Mari C. & Sarah Ogilvie (eds.). 2013. *Keeping languages alive: Documentation, pedagogy, and revitalization*. Cambridge: Cambridge University Press.

Jung, Dagmar & Nikolaus P. Himmelmann. 2011. Retelling data: Working on transcription. Haig, Geoffrey, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.), *Documenting endangered languages: Achievements and perspectives*, 201–220. Berlin: de Gruyter.

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Methods, theory, and practice*. Cambridge: Cambridge University Press.

McEnery, Tony & Nick Ostler. 2000. A new agenda for corpus linguistics – working with all of the world's languages. *Literary and Linguistic Computing* 15(4). 403–420.

Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. *Language Documentation & Conservation* 8. 135–157.

Newman, Paul & Martha Ratliff (eds.). 2001. *Linguistic fieldwork*. Cambridge: Cambridge University Press.

Rice, Sally. 2017. Phraseology and polysynthesism. In Michael Fortescue, Marianne Mithun & Nicholas Evans (eds.), *The Oxford handbook of polysynthesis*, 203–214. Oxford: Oxford University Press.

Rice, Sally & Dorothy Thunder. 2017. Community-based corpus-building: Three case studies. Paper presented at the 3rd International Conference on Language Documentation and Conservation. Honolulu, March 2-5, 2017.

Sakel, Jeanette & Daniel L. Everett. 2012. *Linguistic fieldwork: A student guide*. Cambridge: Cambridge University Press.

Scannell, Kevin P. 2007. The Crúbadán project: Corpus building for under-resourced languages. *Cahiers du Central* 5. 5–15.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Thieberger, Nicholas. (ed.). 2012. *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press.

Thieberger, Nicholas. 2016. Documentary linguistics: Methodological challenges and innovatory responses. *Applied Linguistics* 37 (1). 1–13.

Thieberger, Nicholas & Andrea Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.

Thieberger, Nicholas, Anna Margetts, Stephan Morey & Simon Musgrave. 2015. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21.

Vinogradov, Igor. 2016. Linguistic corpora of understudied languages: Do they make sense? *Káñina* 40(1). 127–141.

Woodbury, Anthony C. 2003. Defining documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description*, vol. 1, 35–51. London: SOAS.

Sally Rice
srice@ualberta.ca
orcid.org/0000-0002-2988-321X